# SPRING 2025: STAT 419 MULTIVARIATE ANALYSIS PROJECT

- **PROJECT GROUPS:** Each STAT 419 student has been assigned to a project group of at most five. See the last page of this document for the group assignments. Each student should exchange contact information with groupmates as soon as possible. Students will have an opportunity to meet groupmates at our next class meeting.
- **DATA ANALYSIS:** Each group has been randomly assigned a data set for the project. See last page for the data set assignments. The data sets (in .CSV format) can be accessed from our Canvas Site (see Course Project). A corresponding PDF describing the data sets can also be found from Canvas.
- To read in CSV files, use the following:
  ```
  dset <- read.csv("C:/DIRECTORY/data.csv", sep=",", header=T)
  ```
- **WHEREVER POSSIBLE, USE RELEVANT FUNCTIONS IN R TO PERFORM THE REQUESTED TASKS. When you do use R, be sure to include the relevant output in the appropriate sections.**

## REPORT FORMAT

Your report should be typed (single spaced, 12 point font), must contain all of the following information, and use the **following section and subsection headings**.

- **COVER:** Project Title, Your Group Number, and All Group Member Names
- **Section A: Introduction/Background**
  Each data set will contain a grouping variable along with other variables. Give a full description of all variables and the purpose of the data set. Although much of that information can be found in the corresponding PDF describing the data set, please provide that information in a more expository manner.
- **Section B: Graphs and Summary Statistics**
  Using R, generate appropriate graphs for the non-grouping variables of the data set. For all quantitative variables, provide a corresponding histogram. When importing graphs into the report, you may want to rescale the graphs as some images can be very large. Be sure to provide comments for each graph (e.g., any interesting patterns?) Also, provide summary statistics of the variables using R. For all quantitative variables, report the corresponding mean, median, and standard deviation.
- **Section C: Discriminant and Classification Analysis**
  Using the methods we discussed in class, conduct a discriminant and classification analysis using the grouping variable and other variables of the data set. Section C must contain the following subsections (be sure to USE THESE SUBSECTION LABELS):
  - **C.1: Correlated Quantitative Variables (Multicollinearity)**
    Provide a table of correlation coefficients for **each pair of quantitative variables** in the data set. This can be given as a correlation matrix that can be generated in R (use the **cor()** command). Also provide the corresponding scatterplot matrix for all quantitative variables. If any two variables appear to have a high correlation (positive or negative) then both variables should NOT be included in your analysis. In that event, describe how you chose which of the two to be included for consideration.
  - **C.2: Discriminant Analysis**
    For this section, be sure to remove any of the correlated variables that were deemed

    unnecessary from Section C.1. It is possible that no variables were found to be sufficiently correlated to be eliminated from consideration. Using the methods described in class, perform the following:

    - Write the complete form of all discriminant functions. Be sure to use standardized coefficients. Based on these coefficients, produce a ranking of variable importance (in the presence of other variables).
    - Carry out tests of significance for the discriminant functions. Be sure to specify corresponding null and alternative hypotheses, test statistic values, and p-values. Be sure to provide a conclusion for each test.
    - Carry out tests of significance of each non-grouping variable, after adjusting the presence of other non-grouping variables. Be sure to specify corresponding null and alternative hypotheses, test statistic values, and p-values. Be sure to provide a conclusion for each test.
    - Produce a plot of the first two linear discriminant functions. Be sure to include this plot in the report. Comment on the plot with regard to how well the discriminant functions separate the groups in the data. NOTE: When using the corresponding function, the system will wait for you to click on the graph to place a legend for the symbols used in the graph. Be sure to click on a location that is empty for the legend placement. If the legend placement causes your system to lock up, modify the original function and remove the code that inserts the legend. In that event, you can manually insert your own legend by annotating the graph.

  - **<u>C.3: Classification Analysis</u>**

    Using the <u>top four most important variables</u> as determined from your discriminant analysis, perform the following tasks:

    - Specify the linear classification functions for each of the group levels in the data.
    - Using Observation #1 from the data, apply the classification functions from the previous step to predict which group that observation should be classified as. Then compare this to the actual group classification from the data and state whether or not the prediction was correct.
    - Based on linear classification functions you generated, provide the corresponding "Confusion Matrix", Apparent Error Rate, and Apparent Correct Classification Rate. Comment on how well the linear classification functions are classifying observations.

- **<u>Section D: Summary</u>**

  Provide a summary of your project. If applicable, include any details that your group encountered (e.g., Any surprising results? Any difficulties arise?) If something did not go as well as you had hoped, describe what you would do the next time. Finally, describe one variable you wish the original data included that your team felt would have been useful for the purpose of group separation.

- **<u>Section E: R Code</u>**

  Provide all the R code you used for your project. Do not include all the corresponding R output. In the previous sections you should have included relevant output from R where appropriate.

## GRADING

Please note that a group can get a very high (or even perfect) score even if their final output does not match my own. In case you haven't noticed, when it comes to data modelling, it's not always true that there's "one unique, sacred, divine, and right answer." Some models are, of course, better than others. But don't fret too much about "have we generated the perfect answer?" because in many situations *there isn't a perfect answer*. Stand by your sensible thought processes and justify your conclusions in a reasonable way. Should you do that, you're probably doing things just fine.

## GROUP MEMBER RESPONSIBILITIES

- **It is important that group members share the responsibility of doing this project.** Any student that fails to sufficiently contribute (thereby causing others to perform a disproportionate amount of work) will face a substantial penalty in project score (**possibly a zero score**). Groups are at liberty to decide how to partition the work. One idea is to have group members write specific portions of the report. But there are many ways to distribute workload.

## PROJECT SUBMISSION DEADLINE: SUNDAY 06/08/25 11:59PM

- Save your project filename in the following format: STAT419GROUP##.PDF. As an example, if your Group number is 5, then the filename will be **STAT419GROUP05.PDF**.
- Submit your typed report (Word, PDF) via CANVAS (DO NOT email to me) **by 11:59pm SUNDAY 6/08/25. The submission should be from only ONE person per group.**
- **DO NOT** send the project as a PDF image scan or as a collection of photographs of your project pages. It must be of a format where I can search the text of your document (for example, your report must be text searchable with the find command [Control-F] or [Command-F]). Aside from checking for statistical accuracy, I'll also check spelling/grammar so please run a spell check before submission.

## PROJECT PEER EVALUATION DEADLINE: TUESDAY 06/09/25 11:59PM

Each STAT 419 student will be required to submit a project peer evaluation. The form will only be seen by me. Your groupmates will not know what you submit. The point of the peer evaluation is for me to know whether or not team members provided sufficient contributions. If any teammate provides insufficient and/or unsatisfactory contributions, it will be important for me to know so that I can assess the appropriate penalty to that student's project score.

The link for this evaluation is here: https://forms.office.com/r/RDyTcvrgRP

You can preview the link now if you would like to see what questions are asked.

**This peer evaluation is due by 11:59pm MONDAY 06/09/25.** If I do not receive an evaluation from **each member** of a group by the deadline, then that group will incur a penalty in project score. Please check with one another to ensure that everyone submits a response by the deadline.

# STAT 419 PROJECT GROUPS

| Student Name | GROUP # |
|---|---|
| Bryukhova, Sonya | 1 |
| Dubow, Max | 1 |
| Jansen, Charlie | 1 |
| Kong, Adam | 1 |
| Tsemekhman, Daniel | 1 |
| Costello, Logan | 2 |
| DeGeorge, Daniel | 2 |
| Kragas, Kailyn | 2 |
| Ramos, Alvaro | 2 |
| Talluri, Sreshta | 2 |
| Cay, Ian | 3 |
| Chu, Karissa | 3 |
| Hegarty, Alexis | 3 |
| Mai, Justin | 3 |
| Vogel, Jack | 3 |
| Drongpa, Abby | 4 |
| Liu, Chris | 4 |
| Palmer, Jett | 4 |
| Sullivan, Alex | 4 |
| Thokala, Sumanth | 4 |
| Brill, Libby | 5 |
| Rajagopalan, Tara | 5 |
| Ralston, Alex | 5 |
| Siegmund, Delaney | 5 |
| Veerasingam, Kaviya | 5 |

| Student Name | GROUP # |
|---|---|
| Hafer, Cameron | 6 |
| Hamilton, Emma | 6 |
| Lamkin, James | 6 |
| Lohier, Anais | 6 |
| Surendran Namboothiri, Jini | 6 |
| Benner, Nate | 7 |
| Dang, Khoa | 7 |
| Hartfelder, Rachel | 7 |
| Kurani, Olivia | 7 |
| Lassa, Owen | 7 |
| Fulton, Lilly | 8 |
| Gamba, Connor | 8 |
| Kerner, Megan | 8 |
| Serrano, Miguel | 8 |
| Solari, Brandon | 8 |
| Garcia, Franchesca | 9 |
| Haruta, Seina | 9 |
| Kuebitz, Charlotte | 9 |
| Manikonda, Adi | 9 |
| Matter, Robin | 9 |
| Chan, Bernette | 10 |
| Degembe, Emi | 10 |
| Harper, Chelsey | 10 |
| Yee, Nicole | 10 |

| GROUP | DATA SET |
|---|---|
| 1 | SAT |
| 2 | cancer |
| 3 | pitchers |
| 4 | cincy |
| 5 | NHANES3 |
| 6 | pollution |
| 7 | cdi |
| 8 | evap |
| 9 | corn |
| 10 | Seishu |